Current Biology, Vol. 13, R901–R903, December 2, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.cub.2003.11.008

Population Genomics: Finding the Variants of Mass Disruption

Dispatch

Greg Gibson

A study of positive selection on the promoter of the human interleukin-4 gene suggests that some disease-promoting alleles might be identified as polymorphic sites that affect transcription and occur at different frequencies among human populations.

Each of us differs from one another by several million snippets of genetic information, and untold millions of life experiences. Out of this complex milieu of variation arise propensities that define our individuality — whether we are likely to suffer from heart disease or depression, explore new worlds or perform great athletic exploits, find contentment raising six kids or discontent working 80 hour weeks. The 'promise' of post-genome human genetics is that we should be able to pinpoint the dozen or so genetic variants that affect any particular predisposition, and as if that is not enough, decipher how and why we differ from our closest primate relatives.

For the past decade or so, the accepted approach to this problem has been some combination of linkage and association mapping [1]. Every year, for hundreds of complex human diseases, several studies are published documenting success or failure in the effort to implicate a candidate gene with the condition. Classical linkage studies are conducted on pedigrees, and basically ask whether affected relatives are more likely to share alleles by descent than are unaffecteds. They can localize a genetic factor to a chromosomal interval, but more resolution is now provided by so-called linkage disequilibrium mapping [2]. This can take the form of population-based case/control studies — basically, is a particular single nucleotide polymorphism (SNP) over-represented among individuals with or without the disease? — or family-based transmission disequilibrium methods — for example, do heterozygous parents transmit both alleles to affected offspring with equal frequency? Sample sizes of several hundred affecteds are required to attain even marginal significance for an association that might explain several percent of the population attributable risk, but replication across multiple studies is required before a finding begins to be taken seriously.

In a paper in this issue of *Current Biology*, Rockman *et al.* [3] argue that population genetic approaches might be used to identify candidate disease-promoting polymorphisms. On the widely accepted assumption that common diseases are caused by common polymorphisms that are nevertheless relatively young in the human gene pool, they suggest that different populations will tend to show different allele frequencies [4] as the species as a whole is not at genetic equilibrium. Their strategy is to search for SNPs that disrupt highly conserved regulatory sequences. If these occur in likely transcription factor binding sites, they then ask whether the SNP frequency varies more than expected by chance among populations.

The case study that Rockman *et al.* [3] report involves a novel binding site for the transcription factor NFAT in the promoter of the interleukin 4 (*IL4*) gene. This site includes a polymorphism at nucleotide –524, where the –524T allele drives more than three-fold greater expression of the IL4 cytokine than the ancestral –524C allele found in the great apes [5]. This variation affects the balance between the two subtypes of T helper cell, Th1 and Th2, in the effector arm of the immune system. This in turn predisposes carriers to chronic allergies, asthma and several infectious respiratory diseases, but probably protects them from other immune challenges including certain forms of HIV and other retroviral infections (see [3] for references). This is an ideal situation for balancing selection to affect disease susceptibility, but as the viral and other environmental challenges undoubtedly vary across populations, the allele frequencies might also vary. Indeed, while the frequency of the –524T allele is 0.45 averaged across the 520 chromosomes surveyed in six populations, it varies from 0.17 in southern Italy to 0.76 in Cameroon (Figure 1).

Rockman *et al.* [3] present two new tests that should convince the reader that this is an unexpected result. First, they compared the differentiation between the six populations with 10,000 random permutations of the frequencies of 18 'neutral' SNPs that they found to be located at least 200 kilobases from any known genes in the human genome. They found that, for five of the fifteen pairwise contrasts of one population against another, the *IL4* site allele frequencies were significantly more divergent than the neutral SNP frequencies. Second, they asked whether the differences were consistent with a model of isolation by distance due to drift, by superimposing the observed distances between populations for IL4 onto a phylogeny derived from the neutral markers. They found that it is highly unlikely that the genetic differentiation at the cytokine locus is consistent with the differentiation observed in the remainder of the genome, and conclude that selection for immune function is the most likely explanation.

How many other examples like this are there likely to be in the human genome? We have no idea, but it is almost certainly more than 'a few'. It is now clear that most, if not all, human genes are polymorphic at the nucleotide sequence level, often including common SNPs located in putative regulatory regions. In fact, Rockman *et al.* [3] began their study with a literature review that grew serendipitously out of control [6]. They started with a couple of papers which documented regulatory SNPs that either affected gene expression in cell

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695-76214, USA.
E-mail: ggibson@unity.ncsu.edu

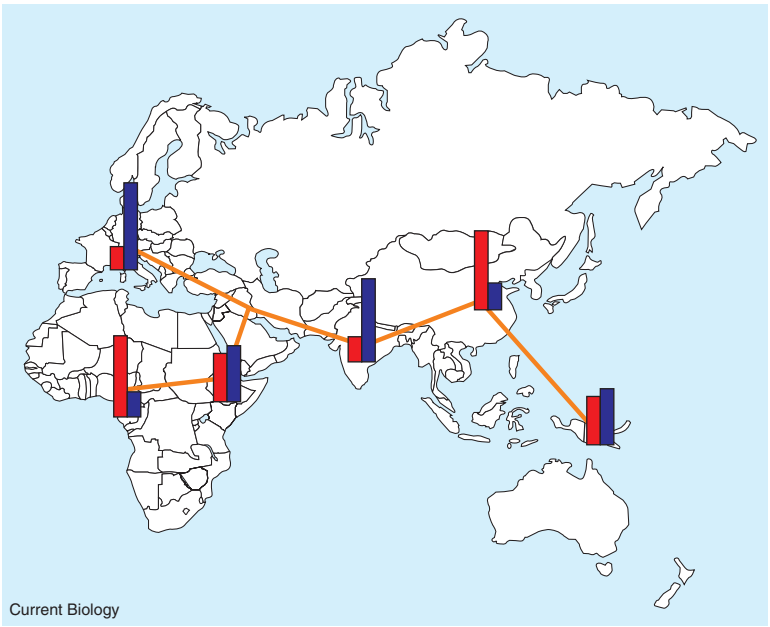

Figure 1. Distribution of *IL4* allele frequencies in the eastern hemisphere.

The frequency of the –524T allele is indicated in red, and that of the –524C allele in blue. The –524T allele is low in Italy and India, intermediate in Ethiopia and New Guinea, and high in Cameroon and China. This distribution does not fit well with the relatedness between the populations defined by genetic distances between neutral markers, implied by the orange lines.

culture experiments or were associated with a clinical or physiological phenotype. Each of these papers referred to another study, and before long Rockman *et al.* [6] had built up a database of more than a hundred such cases, a remarkable finding as in many quarters structural polymorphisms continue to be regarded as the primary source of clinical variation. Yet some reasonable number crunching [6], assuming that the medical genetics literature is relatively unaffected by ascertainment biases, implies that more than half of all human genes may be polymorphic for a site or sites that influence transcription by a factor of twofold or more. That would mean that each of us is heterozygous for as many as 5,000 regulatory SNPs, while the number of functional amino-acid variants is probably smaller than this.

Not only is there abundant raw material for *cis*-regulatory evolution in the human gene pool, but these results also imply abundant possibilities for transcriptional divergence from other mammals. Comparative genomics has been touted as a general approach to identification of regulatory sequences: at least 0.3% of the human genome comprises conserved non-genic sequences [7]. Within these elements, though, individual binding sites are probably much less conserved than previously thought: a recent analysis of experimental and evolutionary data for 51 genes indicated that a striking 32–40% of human functional sites are not functional in rodents [8]. And while it turns out that rodents have evolved at an accelerated rate — dog regulatory regions may be twice as similar to human than are mouse ones [9] — there are literally hundreds of thousands of possible regulatory differences between ourselves and non-primate mammals. For this reason, phylogenetic 'shadowing' — searching for invariant stretches in numerous closely related (monkey) species — holds the most promise for enhancer detection in humans. The approach has already proven useful in locating regulatory sequences in multiple human disease genes [10].

While it seems likely that many genes that exhibit differential expression associated with a disease will turn out to be regulated by *cis*-acting polymorphisms, the corollary will be harder to prove. That is, association of polymorphic functional regulatory sites with disease will continue to be a difficult enterprise. It is not just that the statistical sampling issues are problematic, but the influence of environmental differences on the effects of causative SNPs is now beginning to be investigated. A dietary connection to diabetes has been encapsulated in the 'thrifty genes' hypothesis [11], and altered childhood hygiene may affect allergic responses [12]. More directly, stress and child abuse have been implicated as essential cofactors influencing whether there is an effect of differential expression of the serotonin transporter [13] or the monoamine oxidase A enzyme [14] on depression or juvenile antisocial behavior in a New Zealand cohort. These are very early days for the study of genotype–environment interactions in humans, but the indications are that microarray analyses, in combination with computational promoter analysis, will play an important part identifying candidate genes.

The first gene expression profiling study to compare humans with primates involved a contrast of brain and liver samples from three humans, three chimpanzees and an orangutan [15]. Two independent reanalyses of this experiment concluded that several hundred genes show significant divergence between the species [16,17]. Although there is a suggestion of accelerated divergence in the human brain, perhaps because of cognitive changes — or a slowing down of the divergence between chimp and orangutan, perhaps as a result of growth constraints — in terms of numbers of genes the liver is actually evolving more rapidly than the brain, perhaps reflecting dietary shifts. Most of this transcriptional divergence can probably be explained

by neutral genetic drift [17], but it is nevertheless a profound challenge to identify the genes and sites that make us human, or make us humans sick.

References

1. Hoh, J., and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. Nat. Rev. Genet. *4*, 701–709.
2. Clark, A.G. (2003). Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr. Opin. Genet. Dev. *13*, 296–302.
3. Rockman, M.V., Hahn, M.W., Soranzo, N., Goldstein, D.B., and Wray, G.A. (2003). Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. *13.*
4. Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics *74*, 175–195.
5. Rosenwasser, L.J., and Borish, L. (1997). Genetics of atopy and asthma: the rationale behind promoter-based candidate gene studies. Am. J. Respir. Crit. Care Med. *156*, S152–S155.
6. Rockman, M.V., and Wray, G.A. (2002). Abundant raw material for cis-regulatory evolution in humans. Mol. Biol. Evol. *19*, 1991–2004.
7. Dermitzakis, E.T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., and Antonarakis, S.E. (2003). Evolutionary discrimination of mammalian conserved non-genic sequences. Science *301*, 1033-1035.
8. Dermitzakis, E.T., and Clark, A.G. (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. *19*, 1114–1121.
9. Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M., et al. (2003). The dog genome: Survey sequencing and comparative analysis. Science *301*, 1898–1903.
10. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science *299*, 1391–1394.
11. Diamond, J. (2003). The double puzzle of diabetes. Nature *423*, 599–602.
12. Fuleihan, R.L. (2002). The hygiene hypothesis and atopic disease. Curr. Opin. Pediatr. *14*, 676–677.
13. Caspi, A., Sugden, K., Moffitt, T., Taylor, A., Craig, I., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., et al. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science *301*, 386–389.
14. Caspi, A., McClay, J., Moffitt, T., Mill, J., Martin, J., Craig, I., Taylor, A., and Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. Science *297*, 851–854.
15. Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., et al. (2002). Intra- and interspecific variation in primate gene expression patterns. Science *296*, 340–343.
16. Gu, J., and Gu, X. (2003). Induced gene expression in human brain after the split from chimpanzee. Trends Genet. *19*, 63–65.
17. Hsieh, W.P., Chu, T.M., Wolfinger, R.D., and Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics *165*, 747–757.